

# Toward Order-of-Magnitude Cascade Prediction

Ruocheng Guo , Elham Shaabani, Abhinav Bhatnagar and Paulo Shakarian

Arizona State University

Tempe, AZ

Email: {rguosni, shaabani, abhatn, shak}@asu.edu

**Abstract**—When a piece of information (microblog, photograph, video, link, etc.) starts to spread in a social network, an important question arises: will it spread to “viral” proportions – where “viral” is defined as an order-of-magnitude increase. However, several previous studies have established that cascade size and frequency are related through a power-law - which leads to a severe imbalance in this classification problem. In this paper, we devise a suite of measurements based on “structural diversity” – the variety of social contexts (communities) in which individuals partaking in a given cascade engage. We demonstrate these measures are able to distinguish viral from non-viral cascades, despite the severe imbalance of the data for this problem. Further, we leverage these measurements as features in a classification approach, successfully predicting microblogs that grow from 50 to 500 reposts with precision of 0.69 and recall of 0.52 for the viral class - despite this class comprising under 2% of samples. This significantly outperforms our baseline approach as well as the current state-of-the-art. Our work also demonstrates how we can tradeoff between precision and recall.

## I. INTRODUCTION

When a piece of information (microblog, photograph, video, link, etc.) starts to spread in a social network, an important question arises: will it spread to “viral” proportions – where “viral” is defined as an order-of-magnitude increase. Several previous studies [1], [2] have established that cascade size and frequency are related through a power-law - which leads to a severe imbalance in this classification problem. In this paper, we devise a suite of measurements based on “structural diversity” that are associated with the growth of a viral cascade in a social network. Structural diversity refers to the variety of social contexts in which an individual engages and is typically instantiated (for social networks) as the number of distinct communities represented in an individual’s local neighborhood. Previously, Ugander et al. identified a correlation between structural diversity and influence [?]. We demonstrate these measures are able to distinguish viral from non-viral cascades, despite the severe imbalance of the data for this problem. Further, we leverage these measurements as features in a classification approach, successfully predicting microblogs that grow from 50 to 500 reposts with precision of 0.69 and recall of 0.52 for the viral class (under 2% of the samples).

We note that our results on the prediction of cascades rely solely upon the use of our structural diversity based measures for features and limited temporal features - hence the prediction is based on network topology alone (no content information was utilized). We also achieved these results while maintaining the imbalances of the dataset - which we felt better

mimics reality. This differs from some previous studies which balance the data before conducting classification. Further, we note that we obtained prediction of order-of-magnitude increases in the size of the cascade - which also differs from other work (i.e. [1]) which focus on identifying cascades that double in size.

## II. TECHNICAL PRELIMINARIES

Here we introduce necessary notation and describe our social network data. We represent a social network as a graph  $G = (V, E)$  where  $V$  is the set of vertices and  $E$  as set of directed edges that have sizes  $|V|, |E|$  respectively. The intuition behind edge  $(v, v')$  is that node  $v$  can influence  $v'$ . This intuition stems from how we create the edges in our network:  $(v, v')$  is an edge if during a specified time period there is at least one microblog posted by  $v$  that is reposted by  $v'$  (we leave other thresholds beyond 1 repost to future work). We shall also assume a partition over nodes that specifies a community structure. We shall assume that such a partition is static (based on the same time period from which the edges were derived) and that the partition  $\mathcal{C}$  consists of  $k$  communities:  $\{C_1, C_2, \dots, C_k\}$ . There are many possible methods to derive the communities (if user-reported communities are not available). We utilize the Louvain algorithm to identify our communities in this paper due to its ability to scale.

**Cascades.** For a given microblog  $\theta$ , we denote the subset of first- $m$  nodes who originally posted or reposted  $\theta$  as  $V_\theta^m$  and refer to them as *adopters* (at size  $m$ ). Likewise, the set of reposting relationships within the same time period will be denoted  $R_\theta^m$ . Taken together, we have a *cascade*:  $D_\theta^m = (V_\theta^m, R_\theta^m)$ . Any valid original microblog  $\theta$  could be treated as a unique identifier for a cascade. Given a microblog  $\theta$ ,  $v_\theta$  is the originator at instance  $t_\theta^0$ , which is defined as the origin time when the originator posted the microblog  $\theta$  and time  $t$  is time since  $t_\theta^0$ . The  $m^{th}$  repost of the microblog  $\theta$  happens at time  $t_\theta^m$ . As  $m$  increases, a cascade accumulates nodes and edges over time. We shall use  $N_\theta$  to denote the final size of a cascade while the size of a cascade at any particular instance is the set of nodes present at that instance is simply  $|V_\theta^m|$ . For a given size  $m$ , we shall refer to the *frontiers* as the outgoing neighbors of the adopters in graph  $G$  who are not adopters themselves. Formally:  $F_\theta^m = \{v \in V/V_\theta^m \text{ s.t. } \exists v_i \in V_\theta^m \text{ where } (v_i, v) \in E\}$ . For nodes in  $G$  that are outside the adopters, we shall use the notation  $t_{exp}(v, \theta, m)$  to denote the number of time units from the initial post of  $\theta$  before the microblog was reposted by one of  $v$ ’s incoming neighbors - intuitively the time at which  $v$  was exposed to  $\theta$ . For a given natural number  $\lambda$  (used to specify a time period), we define the  $\lambda$  frontiers as a subset of the frontiers that have been exposed to  $\theta$

U.S. Provisional Patent 62/201,517. Contact shak@asu.edu for licensing information

no earlier than  $\lambda$  time units previously. Formally this set is defined as follows:  $F_{\theta}^{m,\lambda} = \{v \in F_{\theta}^m | t_{exp}(v, \theta, m) \leq \lambda\}$ . Finally, the complement of this set are the  $\lambda$  non-adopters:  $\bar{F}_{\theta}^{m,\lambda} = \{v \in F_{\theta}^m | t_{exp}(v, \theta, m) > \lambda\}$ .

**Sina Weibo Dataset.** The dataset we used was provided by WISE 2012 Challenge<sup>1</sup>. It included a sample of microblogs posted on Sina Weibo from 2009 to 2012. In this dataset, we are provided with time and user information for each post and subsequent repost which enabled us to derive a corpus of cascades. From this data, we derived our social network  $G = (V, E)$  (with 17.9 M vertices and 52.4 M edges) that was created from reposts that were published during the 3 month period between May 1, 2011 and July 31, 2011. For this network, the average clustering coefficient is 0.107. There are 4974 connected components in the network. Louvain algorithm outputs 379,416 communities with average size of 47.5 for this network. As expected, this network exhibits a power-law degree distribution. For this network, the number of active nodes in August (the time period we studied for cascade prediction) is 5,910,608, while 5,664,625 of them at least have one out-neighbor. During the month of August, there were 9,323,294 reposts with 2,252,368 different original microblogs. 1,920,763 (86.6%) of them were written by authors who at least published one microblog during May 1, 2011 to July 31, 2011 (the time period we used to create the underlying network). The average time it took for viral cascades to become viral is approximately 18 hours. The distribution of final size of cascades mimics a power-law distribution which can demonstrate that this dataset is more representative of cascade behavior observed “in the wild”. This differs significantly from the previous works which conduct biased sampling to artificially provide balanced classes. We selected  $\lambda$  as 30 minutes as 90% of all reposts in the initial 3 month period occurred in under this time.

**Number of communities.** For  $V' \subseteq V$ , the associated communities  $\mathcal{C}(V')$  are the communities represented by  $V'$ . Formally:  $\mathcal{C}(V') = \{C_i \in \mathcal{C} \text{ s.t. } V' \cap C_i \neq \emptyset\}$ . The cardinality of this set (number of communities) will be denoted  $K(V')$ . We measure the number of communities represented by the above three populations of nodes:  $K(V_{\theta}^m)$ ,  $K(F_{\theta}^{m,\lambda})$ ,  $K(\bar{F}_{\theta}^{m,\lambda})$  observed at either a given cascade size.

**Gini impurity.** For  $V' \subseteq V$ , the gini impurity,  $I_G(V')$  is the probability of a node in  $V'$  being placed into the incorrect community if assigned a community based on the distribution of communities represented in  $V'$ . Formally:  $I_G(V') =$

$1 - \sum_i (\frac{|C_i \cap V'|}{|V'|})^2$ . We study the gini impurity of the adopters,  $\lambda$  non-adopters and  $\lambda$  frontiers for either a given cascade size  $m$ :  $I_G(V_{\theta}^m)$ ,  $I_G(F_{\theta}^{m,\lambda})$ ,  $I_G(\bar{F}_{\theta}^{m,\lambda})$ . The intuition is to capture a notion of how the communities are distributed amongst the nodes in each of these sets with a single scalar value. We note that the impurity of the adopter set  $I_G(V_{\theta}^m)$  behaves similar to the entropy of this set (a measurement introduced in [3]). However, as we will see in the next two sections, we found that the impurity of the  $\lambda$  frontiers is a more discriminating feature.

**Overlap.** For  $V_a, V_b \subset V$ , the overlap ( $O(V_a, V_b)$ ) is simply the number of shared communities. Formally:  $O(V_a, V_b) =$

$|\mathcal{C}(V_a) \cap \mathcal{C}(V_b)|$ . We study overlap between adopters and  $\lambda$  frontiers, between adopters and  $\lambda$  non-adopters, and between  $\lambda$  frontiers and  $\lambda$  non-adopters:  $O(V_{\theta}^m, F_{\theta}^{m,\lambda})$ ,  $O(V_{\theta}^m, \bar{F}_{\theta}^{m,\lambda})$ , and  $O(F_{\theta}^{m,\lambda}, \bar{F}_{\theta}^{m,\lambda})$  respectively. The intuition with overlap stems directly from the original structural diversity results of [?] - for instance a high overlap between adopters and  $\lambda$  frontiers may indicate that the  $\lambda$  frontiers are linked to adopters with inner-community connections and high structural diversity - hence increasing the probability of adoption.

**Average time to adoption.** The average time to adoption for the nodes in the current set of adopters (once the cascade grows to size  $m$ ):  $\frac{\sum_{i=1}^m t_{\theta}^i}{m}$ . We also use average time to adoption as a baseline measure.

### III. RESULTS

Here we examine the behavior of the various structural diversity measurements as viral and non-viral cascades progress. We define a cascade as viral if the number of reposts reaches a threshold (denoted  $TH$ ) of 500 (in the next section we will explore other settings for  $TH$  when describing our classification results). We look at snapshots of the cascades as they progress both in terms of size (denoted  $m$ ). For  $m = \{10, 30, 50, 100, 200\}$ , the number of samples is  $\{98832, 26733, 13285, 4722, 1324\}$  respectively with 208 of the samples are viral. With each size  $m$  we consider the Cascades with  $m$  adopters at some time  $t_{\theta}^m$ ,  $t_{\theta}^m$  can vary for different  $\theta$ . Hence, cascades with final size  $N < m$  are ignored in our analysis task. This leads to a decrease in the number of non-viral Cascades as  $m$  increases.

**Average time to adoption.** As a baseline measurement, we study the average time to adoption for each size-based stage of the cascade process (Fig. 1i, Fig. 1j). As expected, viral cascades exhibit a faster rate of reposting. While we note that significant differences are present - especially in the early stages of the cascade, the whiskers of the non-viral class indicate a significant proportion of non-viral cascades that exhibit rapid adoption. We believe this is likely due to the fact that certain cascades may have very high appeal to specialized communities.

**Number of communities.** Fig. 1a, Fig. 1b, Fig. 1c and Fig. 1d display how the number of communities  $K(V')$  increases over  $m = \{10, 30, 50, 100, 200\}$  for the sets  $V' = \{V_{\theta}^m, F_{\theta}^{m,\lambda}\}$ . We note that  $K(V_{\theta}^m)$  (the communities represented in the set of adopters) was shown to be a useful feature in [3] for tasks where the target class had fewer reposts than in this study. Here, we note that while statistically significant differences exist, the average and median values at each of the examined stages are generally similar. On the other hand, the communities represented by the set of  $\lambda$  frontiers ( $K(F_{\theta}^{m,\lambda})$ ) shows viral Cascades have stronger capability than non-viral ones to keep a diverse set of  $\lambda$  frontiers. We also noted that the median of  $K(\bar{F}_{\theta}^{m,\lambda})$  (not pictured) shows viral cascades start with smaller  $K(F_{\theta}^{m,\lambda})$ . However, it increases faster in viral cascades as nodes in  $\lambda$  frontiers becomes  $\lambda$  non-adopters.

**Gini impurity.** Cascades in both classes tend to accumulate diversity in the process of collecting more adopters - and we have also noted that a related entropy measure (studied in [3]) performed similarly. We also noted (not pictured) that in the

<sup>1</sup><http://www.wise2012.cs.ucy.ac.cy/challenge.html>

TABLE I: Features: Cascade Prediction over Time and Size

Group	Feature(s) over size
$A_m$	$K(F_\theta^{m,\lambda}), K(\bar{F}_\theta^{m,\lambda}), I_G(V_\theta^m), I_G(F_\theta^{m,\lambda}), I_G(\bar{F}_\theta^{m,\lambda}),$ $O(V_\theta^m, F_\theta^{m,\lambda}), O(V_\theta^m, \bar{F}_\theta^{m,\lambda}), O(F_\theta^{m,\lambda}, \bar{F}_\theta^{m,\lambda}),$ $ F_\theta^{m,\lambda} ,  \bar{F}_\theta^{m,\lambda} , \frac{\sum_{i=1}^m t_\theta^i}{m}, m \in \{30, 50\}$
$B_m$	Community Features Mentioned in [3] and $C_m$
$C_m$	$\frac{\sum_{i=1}^m t_\theta^i}{m}, m = 50$

early stages, viral cascades can show more diversity in  $\lambda$  frontiers measured by  $I_G(F_\theta^{m,\lambda})$  ( $m = \{10, 30, 50\}$ ). But, perhaps most striking, that non-viral Cascades gain more uniformly distributed nodes over communities in  $\lambda$  non-adopters, shown by  $I_G(\bar{F}_\theta^{m,\lambda})$  (Fig. 1g, Fig. 1h). We believe that this is due to non-viral cascades likely have an appeal limited to a relatively small number of communities - hence those *not* adopting the trend may represent a more diverse set of communities.

**Overlap.** We found that overlap grows with the number of adopters in the three types of overlap considered. For  $O(V_\theta^m, F_\theta^{m,\lambda})$ , viral cascades start with a larger initial value and keep leading non-viral ones in the diffusion process of first 200 nodes (Fig. 1e, Fig. 1f). This may hint that viral cascades also take advantage of the densely linked communities to help them become viral. However, in the case of  $O(V_\theta^m, \bar{F}_\theta^{m,\lambda})$  and  $O(F_\theta^{m,\lambda}, \bar{F}_\theta^{m,\lambda})$ , viral cascades begin with lower value but grow much faster than non-viral Cascades.

**Classification Experiments.** Here we examine our experiments for predicting whether a cascade becomes viral - when a size threshold ( $TH$ ) exceeds 500 adopters given that the cascade has 50 adopters ( $s = 50$ ). Based on the distribution of final size of cascades in this dataset, this is a binary classification task with two heavily imbalanced classes. *Hence, we report performance measurements (precision, recall and F1 score) for only the minority (viral) class.* Throughout the course of our experiments, we found that varying threshold (slightly modifying the definition of “viral”) for *only* the training set allows for a trade-off between precision and recall. We study the trend of performance measures in two cases: (1.) The threshold for test set is maintained as  $TH_{ts} = 500$  while the training threshold is varied  $TH_{tr} = \{300, 400, 500, 600, 700\}$ . (2.) The two thresholds are kept as the same  $TH$  while we modify this value  $TH = \{300, 400, 500, 600, 700\}$ .

Table I shows the groups of features used in our prediction tasks. The features introduced in this paper is group  $A_m$ . As a baseline method for size-based prediction (feature group  $C_m$ ) we used average time to adoption. We also compare our features (Group  $A_m$ ) with the community features extracted in [3] (Group  $B_m$ ). This was the best performing feature set in that paper for a comparable task.<sup>2</sup> Additionally, we study the average size of recalled and non-recalled viral cascades by classifiers using features in groups  $A_m$ . We also investigate the significance and performance of individual and certain combinations of features introduced in this paper.

We used ten-fold cross-validation in our experiments to

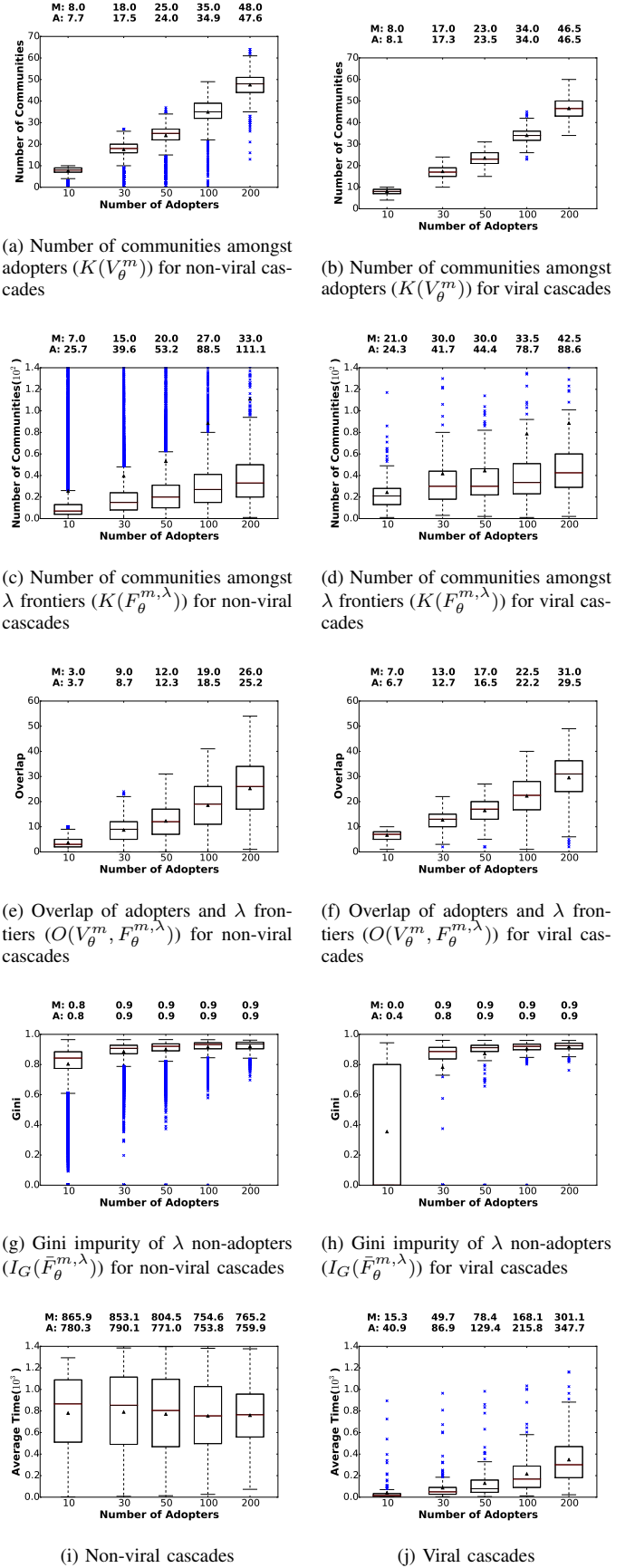


Fig. 1: Number of communities, gini impurity, overlap and average time since  $t_0^0$  to adoption for  $m = \{10, 30, 50, 100, 200\}$

<sup>2</sup>This was their highest-performing set of features for predicting cascades that grew from 50 to 367 and 100 to 417 reposts. We also included the baseline feature in this set as we found it improved the effectiveness of this approach.

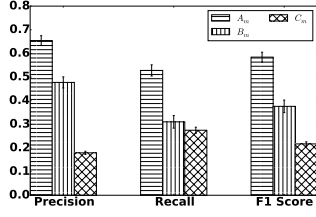
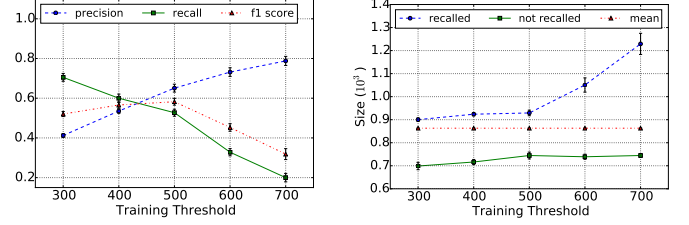


Fig. 2: Classification results based on groups of features ( $A_m, B_m, C_m$ ) extracted when  $m = 50$  for fixed  $TH_{tr} = 500$ ,  $TH_{ts} = 500$ . Error bars represent one standard deviation.

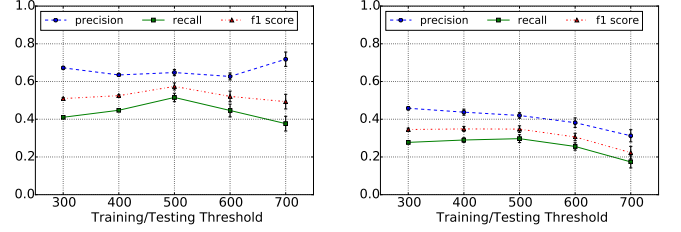
ensure the results do not take any advantage of randomness in picking training and testing sets. First we carried out the prediction tasks with fixed thresholds  $TH_{tr} = 500$ ,  $TH_{ts} = 500$ . Then we modify the training threshold  $TH_{tr} = \{300, 400, 500, 600, 700\}$  to show how this achieves a trade-off between precision and recall. The difference in average final size between correctly classified viral cascades and incorrectly classified ones is also monitored over  $TH_{tr} = \{300, 400, 500, 600, 700\}$  to show the potential to predict exact number of adopters by features. Furthermore, we modify threshold of both training and testing sets  $TH = \{300, 400, 500, 600, 700\}$  to show the robustness of our features on related classification problems. We used the oversampling method SMOTE with random forest classifier to generate synthetic samples for the viral class. Other, lesser-performing classifiers were also examined (including SVM, MLP, and other ensemble methods) and are not reported here. All results shown in this section is a sample mean produced by ten repeated experiments under each combination of variables.

**Size-based prediction.** We studied cascades of size 50 that reached 500 for this task. There are 13,285 cascades that can reach the size  $m = 50$  while 208 out of them reached the size of 500. Maintaining the threshold  $TH = 500$ , Fig. 2 shows random forest classifier trained with features in group  $A_m$  can outperform the other groups. The trade-off between precision and recall can be achieved by changing the training threshold  $TH_{tr}$  while maintaining the testing threshold (Fig. 3a). We also note that the average final size of viral cascades recalled by the classifier increases with the training threshold (Fig. 3b). With threshold  $TH = \{300, 400, 500, 600, 700\}$  on both training and testing samples, the features of group  $A_m$  consistently outperform those previously introduced ( $B_m$ ) (Fig. 3c, Fig. 3d).

**Feature investigation.** Here we investigate the importance of each feature in  $A_m$ . With  $TH_{tr} = 500$  and  $TH_{ts} = 500$ , we trained 100 randomized logistic regressions models - each assigning weights to the features in those sets. We then categorized the features with weight larger than 0.01 (on average) into groups such as overlap, gini impurity, etc. Then, we performed classification on the basis of single feature categories or combination of such categories. The average weights assigned are shown in Table II. As shown by these results, overlaps can make significant contribution to cascade prediction. Intuitively, communication between two sets of nodes is more likely to happen in their shared communities - which is consistent with the results of [?]. This implies that the larger overlap value, the more influence of one set on the other. For example, we can infer that viral cascades tend to have



(a) Results for features in  $A_m$  with different  $TH_{tr}$ . (b) Average final size of viral cascades (recalled, mean and not recalled)



(c) Results for features in group  $A_m$  when  $TH_{tr}$  and  $TH_{ts}$  change. (d) Results for features in group  $B_m$  when  $TH_{tr}$  and  $TH_{ts}$  change.

Fig. 3: Prediction results for  $A_m$  when  $TH_{tr}$  and  $TH_{ts}$  change. Error bars represent one standard deviation.

Name	Features	Weights
Gini	$I_G(F_{\theta}^{50, \lambda})$	0.02
	$I_G(\bar{F}_{\theta}^{50, \lambda})$	0.02
	$I_G(\bar{F}_{\theta}^{30, \lambda})$	0.52
Baseline	$\frac{\sum_{i=1}^{50} \tau_i}{50}$	1.00

Name	Features	Weights
Overlap	$O(V_{\theta}^{30}, \bar{F}_{\theta}^{30, \lambda})$	0.50
	$O(V_{\theta}^{30}, \bar{F}_{\theta}^{50, \lambda})$	0.04
	$O(F_{\theta}^{30, \lambda}, \bar{F}_{\theta}^{30, \lambda})$	0.23
	$O(V_{\theta}^{50}, \bar{F}_{\theta}^{50, \lambda})$	0.50
	$O(F_{\theta}^{50, \lambda}, \bar{F}_{\theta}^{50, \lambda})$	0.26

TABLE II: Weights of features assigned by randomized logistic regression models

larger  $O(V_{\theta}^m, F_{\theta}^{m, \lambda})$  value for adopters have larger chance to influence the  $\lambda$  frontiers than non-viral cascades. Moreover, the gini impurity of  $\lambda$  non-adopters also shows its importance. Intuitively, non-viral cascades are easier to be trapped in a relatively small amount of communities. This means even if they could show up in people's timeline with high structural diversity but can not get them infected.

#### IV. ACKNOWLEDGMENT

This work is supported through the AFOSR Young Investigator Program (YIP), grant number FA9550-15-1-0159.

#### REFERENCES

- [1] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee, 2014, pp. 925–936.
- [2] P. Shakaran, A. Bhatnagar, A. Aleali, R. Guo, and E. Shaabani, *Diffusion in Social Networks*. Springer (in press), 2015. [Online]. Available: <http://lab.engineering.asu.edu/cysis/diffusion/>
- [3] L. Weng, F. Menczer, and Y.-Y. Ahn, "Predicting successful memes using network and community structure," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.